Group 3 Rough Draft

Benjamin Madnick* Nolan Karsok* Noah DCruz Naveem Nanda Mohan
Department of Mathematics, Northeastern University

December 11, 2022

Abstract

Classification is a supervised machine learning approach where the result is that the data has categorized into several classes. These classes can be binomial, such as a positive or negative classification, or multi-class. The different types of cancer is an example of a multi-class classification problem. Another useful machine learning technique is feature importance which determines the most important features for a model. This reduces the dimensionality of a problem and simplifies the complexity of the model.

Our goal is to learn which machine learning model introduced yields the most accurate prediction for each type of cancer, and to identify which genes are the most important in relation to gene effect, expression, and dependence.

1 Introduction

As medicine and technology reach unprecedented levels of advancement, there are still a number of diseases that scientists and doctors are trying to combat. These diseases include Alzheimer's, Parkinson's, Huntington's, and the various types of cancer that plague different parts of the body. As the population and life expectancy increase, so does the number of appearances of cancer. As a result, oncologists are approaching cancer identification and expression through the lens of data science.

In this paper, we present a method that runs various machine learning models and selects the best model for each type of cancer, based on results such as accuracy, precision score, f1 score, and more. In addition, we implement a feature importance algorithm and machine learning model in order to determine which genes have the greatest correlation to each different type of cancer.

2 Related work

Investigation of Gene-Cancer Associations A 2020 Scientific Reports study, "Genome-wide investigation of gene-cancer associations for the prediction of novel therapeutic targets in oncology", conducted by Adrián Bazaga, Dan Leggate and Hendrik Weisser of Storm Therapeutics, focused on a similar goal of identifying differential cancer gene dependencies and classifying predictive models sharing genetic determinants using CRISPR KO and mutational datasets. The authors of this paper analyzed the performance of the following machine learning classifiers: random forests, artificial neural networks, support vector machines, logistic regression and gradient boosting. These models were trained on three features that were specific to each individual cancer type: gene-expression, mutation and essentiality. The model performance in this study varied across the different cancer types, where the highest AUROC scores were close to 0.90 and the lowest AUROC of 0.75, coming from the leukemia model. The poor performance of the leukemia model was explained by the large number of leukemia subtypes, all of which were combined in this dataset. The following case study looks to reproduce similar results of this Scientific Reports paper across the same datasets.

3 Formulation

3.1 Data Generation

The datasets that were used in this study consisted of the publicly available genomic datasets. There were two main technologies used to generate this dataset: clustered regularly interspaced short palindromic repeats (CRISPR) and Knowckout (KO). The CRISPR technology is used by research scientists to modify the DNA of living organisms and KO is used to inactivate or remove specific genes from an organism. A series of pre-processed gene-level expression data that utilized this technology were downloaded from depmap.org and transformed, and a description of each dataset is below.

CRISPR Gene Effect Using the CRISPR Knockout technology, this dataset consists of Gene Effects Scores that were originally published by Broad's Achillies and Sanger's SCORE projects. To interpret this data, negative scores represent cell growth and inhibition (or death) following gene knockout.

CRISPR Gene Dependency This dataset contains probabilities representing the likelihood that knocking out or removing the gene has a growth inhibition or death effect.

CCLE Expression The CCLE Expression dataset contains transcript per million (TPM) values of protein coding genes. This data is preprocessed after log2 transformation and uses a pseudo count of 1

CCLE Mutations This dataset contains metadata that is used in the other mutations dataset. It contains various information on the somatic point mutations that are referenced in the DepMap cell lines.

CCLE Gene CN This is another preprocessed dataset that is copy number data at the gene-level and also log2 transformed with a pseudo count of 1. Additional copy number datasets can be downloaded on depmap.org but were not used for this analysis.

CCLE WES Gene CN This is another preprocessed dataset that is copy number data at the gene-level and also log2 transformed with a pseudo count of 1 and is inferred from the WES data which maps genes onto the segment level calls

Sample Info This final dataset is the metadata for DepMap's cancer models and cells. The lineage field in this table represents the standardized cancer type classifications which is our target variable.

3.2 Classification

Using the CRISPR Gene Effect, CCLE Expression, and CRISPR Gene Dependency datasets, we began to classify which machine learning models are the best at identifying certain types of cancer. The primary focus was on the following cancer types: breast, colon, kidney, leukemia, lung, liver, ovarian, and pancreatic cancer. In order to classify cancer lineages across patients, each dataset (effect, expression, dependency) was joined onto the sample info dataset. From there our target data was the lineages from the sample info dataset and the features were the genes from each data set. Using these three datasets we were able to classify the lineages.

For each model a classification report is generated that contains the precision, f1 score, and recall for each cancer type. The average precision across all the machine learning models is compiled below in the results section.

3.3 Feature Importance

In order to determine which genes were the most importance, a random forest model was implemented. Feature selection is useful in reducing the dimensionality and complexity of a machine learning model. By isolating which features are important and dropping the rest, the results of a classification method can be improved. Using the datasets from DepMap, we can generate an intersection of which features are important across all of the datasets. Using these to train the machine learning models should produce results that are improved. Besides improving models, feature selection can also be useful in identifying which genes should be removed using CRISPR KO. By determining an importance threshold, only those genes that have a certain importance can be expected to be knocked out.

4 Implementation

Linear Discriminant Analysis The first model that was implemented was linear discriminant analysis (LDA). A benefit of the LDA method is it is quick and simple to produce while still maintaining strong performance. LDA models are often used in medical practices to classify patient diseases and can be used by medical professionals to offer the correct treatment given the model parameters.

Quadratic Discriminant Analysis We attempt to run a quadratic discriminant analysis model on the dataset. Using QDA we had to drop certain cancer types because the number of values was too low, and this was causing issues with the covariance calculations. QDA is good to use in conjunction with LDA because it uses more parameters than LDA which can lead to better results.

Logistic Regression We implement a logistic regression model which is a good multiclass classification technique. When running a logistic regression model it is worth paying attention to the precision/recall trade-off. Precision is the proportion of positive identifications that was actually correct while recall is the proportion of actual positives that was identified correctly. When it comes to the analysis of the logistic regression, this is the focal point.

Support Vector Machine Support Vector Machine is primarily for classification although it can be customized for regression tasks. SVM is a supervised machine learning model which segregates data across an n-dimension hyperspace using a boundary line known as a hyperplane. The hyperplane is generated using "support vectors" (Points at the boundary of each class) and a kernel function (A function that applies transformation to convert lower order data to higher dimensions).

For our purposes, we are implementing SVM with Soft margin. Soft margin is a technique in which the model is given a wider gap to misclassify data points by setting a flexible margin of error. By doing so, performance of the model can be enhanced and non-linearly separable data can be handled.

With SVM there are certain parameters that have to be tuned. One parameter that has to be tuned in C, which is a regularization parameter used to define the margins, with soft margin $C = 1.0$. The kernel, which is used for applying higher order transformations, was selected to be the linear kernel. The degree of the polynomial expression was also tuned, as was the kernel coefficient γ, which determined the curvature of the decision boundary.

Random Forest The random forest model is a non parametric model that can be used for both classification and regression. Random forest can also be used for feature selection and feature importance. In our case we used the RF method for feature importance.

Using a split between test and train data, the random forest model is trained and fit in the same manner as the other models implemented in this paper, but instead of a classification report we return a ranked list of genes with their relative importance. From there we can find a true feature importance list which is just the intersection of each feature importance list for each model.

5 Results

5.0.1 Linear Discriminant Analysis

The first implemented LDA model on the CCLE Expression dataset showed a relatively strong performance where it successfully classified 78% of cancer type lineages across patients using the gene expression transcript per million (TPM) values of protein coding genes from the CCLE Expression dataset. The cancer type with the highest value count of patients (14% of all records were of lung cancer lineage) had one of the lowest model precisions of only 53%. However, lung cancer had a higher recall in the LDA model of 85%, meaning that this portion of the model had a high false positive rate. Generally, in practices of classifying cancer types using machine learning algorithms similar to the work in this paper, diverse cancers are often more difficult to predict. Cancer types such as blood (leukemia), lung, thyroid, etc. all have a high number of subtypes and classifying them all under the same umbrella can cause poor model performance. This is consistent in what was seen with this LDA model, especially in the lung cancer lineage and despite having a high number of datapoints, the model still had a low precision. A future study of the LDA model could be separating out only the lung cancer lineage and further breaking it down into different subtypes, and using those subtypes with the gene expression data to classify the more granular subtype of lung cancer to yield of higher precision and reduce the false positive rate.

The second LDA model that was implemented using all features was around the CRISPR Gene Effect dataset and again, predicting cancer type lineage. This was the second best performing model as it successfully classified 61% of cancer type lineages using the features from the CRISPR Gene Effect dataset. Similar to the model using CCLE Expression data, lung cancer was among the lowest performing cancer types with a precision score of only 0.27. The recall score remained high at 0.86.

Lastly, the third dataset that an LDA model was implemented on, with an overall precision score of only 0.59 was the CRISPR Gene Dependency dataset. This model performed lower than the previous two across many cancer types. The probability of knocking out the gene has a cell growth inhibition from the CRISPR Gene Dependency data has proved to not be very predictive of cancer type lineage.

5.0.2 Quadratic Discriminant Analysis

For our QDA models, we obtained an average precision of 0.08 from the 3 datasets. For the gene expressions dataset, the average precision was 0.11. In addition, liver and colon had a precision score of 0.00. Leukemia had the highest precision at 0.24, followed by lung at 0.16 and ovarian at 0.13.

For the gene dependency dataset, the average precision was 0.06. The following cancers had a precision score of zero: liver and ovarian. Lung had the highest precision at 0.22.

For the gene effect dataset, the average for precision was 0.07. The following cancers had a precision score of zero: breast and liver. Similar to the gene expression dataset, leukemia had the highest precision score at 0.24, followed by lung at 0.16 and ovarian at 0.13. It is worth noting that liver had a precision score of 0.00 across all three datasets using a QDA model. This could be due to some intrinsic qualities of the liver cancer data, or could be due to a lack of sufficient data points for that lineage type.

Overall, QDA is a poor model to run on the data and did much worse in comparison to the LDA method. This could be due to the data as LDA is a much less flexible classifier than QDA, thus it a has substantially lower variance. Meaning that LDA does better than QDA in terms of linearity. Another possible explanation for why the QDA results were worse is due to the non-normality of the datasets.

5.0.3 Logistic Regression

Overall the logistic regression had an average precision of 68% with an average recall of 66%. Thus when we consider the precision-recall trade-off, we would prefer a higher

recall because that means there were less false negatives. This is ideal over a higher precision because we would want more people to be correctly identified with the type of cancer present.

First when dealing with the gene expression dataset, we found a precision of 77% and recall of 76%. Liver cancer had a precision value of 1.0 with this data, so there were no false positives. It is also worth noting a recall value of 1.0 for kidney and liver cancer using this data, which indicates that for those cancer types there were no false negatives. This dataset also yielded the highest precision and recall overall among the three datasets tested.

LR was then performed on the gene effect dataset, which had an average precision of 62% and average recall of 63%. The main statistic of note with this dataset was the 0 precision and recall for liver cancer. This could be due to only there only be a 1.8% sample of the lineages being liver cancer. Or it could be due to complexities in something intrinsic to liver cancer.

The final dataset was gene dependency. This was the worst dataset, as the precision was only 64% and the recall 60%. An interested result of this model, however, was the precision value of 1.0 for liver cancer, while the recall was at .33. So when looking at liver cancer and logistic regression it is better to prioritize this dataset and gene expression in order to get a better classification result.

5.0.4 Support Vector Machine

From the distribution of the classes, we can see the disparity in the amount of training data available for each class. This is an important cause for the different scores across each cancer type. A preliminary analysis was performed using a custom SVM algorithm, with underwhelming results. This is not new as SVM is known to perform poorly for datasets with a lot of features. The default preferable kernel used is the RBF, which does not do well with linear data. When setting the kernel to linear, the results jump by a substantial margin due to the linear separable nature of the data. When comparing the results of the other models to SVM, SVM has an edge due to the implemented kernel and the soft margin implemented through the C hyper parameter. This means that certain data points are a poor representative of the over all dataset, which implores the need for k fold cross validation. Bias variance trade-off was handled by creating an unfamiliar test case to establish the wide gap of data points not covered by soft margin. Huber loss turns out to be the most optimal solution as linear loss is covered due to the offset of bias.

5.0.5 Comparison against established models

Looking at the published cancer model performance data, we can make a comparison based on the values associated with each model under each type of cancer

cancer-wide_models_performance

Model \ Cancer type	Bladder	Breast	Colon	Kidney	Leukemia	Liver	Lung	Ovarian	Pancreatic
Logistic regression	0.78	0.77	0.69	0.86	0.75	0.84	0.81	0.79	0.75
Support Vector Machine	0.77	0.78	0.72	0.88	0.72	0.84	0.87	0.8	0.73
Gradient Boosting Machine	0.75	0.7	0.74	0.75	0.71	0.81	0.73	0.74	0.73
Neural Network	0.67	0.72	0.71	0.71	0.7	0.86	0.75	0.75	0.72
Random Forest	0.76	0.75	0.76	0.79	0.74	0.85	0.83	0.77	0.76

Figure 1: Established model performance for each cancer type

Model/Cancer type	Breast	Colon	Kidney	Leukemia
LDA	.82	.89	.86	.98
QDA	.05	.06	.08	.18
LR	.70	.73	.79	.83
SVM	.83	.89	.81	.98

Model/Cancer type	Liver	Lung	Ovarian	Pancreatic
LDA	.27	.35	.47	.52
QDA	0.00	0.18	.09	.18
LR	.67	.57	.68	.75
SVM	.33	.42	.54	.68

5.0.6 Feature Importance

When we run the random forest model for each dataset, we utilize the build in feature importance which is based on mean decrease in impurity. Given the top 100 features for each dataset, we obtain an intersection which contains the importance features across all three datasets.

	importance	importance	importance
SOX10 (6663)	0.001385	0.001485	0.001489
PAX8 (7849)	0.001342	0.002749	0.001730
CDX2 (1045)	0.000997	0.000934	0.000389
MYB (4602)	0.000984	0.001119	0.002612
DOCK2 (1794)	0.000864	0.000746	0.001040

Figure 2: Intersection of the feature importance data frame across all three datasets

We can see this intersection only contains five genes. When it comes to CRISPR KO, these should be the genes that should be removed. Considering there are 19,000 genes in each dataset, it is interesting that there are only five that have a significant common importance across all three datasets. However with only five features, running the models again might not yield beneficial results. Especially considering each gene only has a small fraction of importance due to the number of genes that the dataset contains. A possible future step would by the run all the models again using the top 100 genes for each dataset.

6 Acknowledgement

We would like to thank Dr. Suraj Peri, and the rest of the team at Merck Research Laboratories and Prof. He Wang for collaborating and making this XN project viable for the students in this class. This project introduced the members of this group to various skills that will be useful in the future, such as working with real data and referencing academic papers. Additionally, this project acted as a networking opportunity due to its association with Merck Laboratories, which is beneficial to MS students who will soon be entering into industry.

References

[1] Adrian Bazaga, Dan Leggate, and Hendrick Weisser, *Genome-wide investigation of gene-cancer associations for the prediction of novel therapeutic targets in oncology*, Scientific Reports (June 1, 2020)